基于用户特征属性的微博话题关键用户挖掘 *

柯　阳，隋　杰†

(中国科学院大学 工程科学学院，北京 100049)

摘　要：针对微博话题存在时效性的特征以及用户之间交互行为特征，在经典 PageRank 算法的基础上，提出基于用户交互的微博用户挖掘算法来有效挖掘推动微博话题流行的关键用户。首先，介绍了微博话题关键用户的定义及其相关特征；其次，由于传统模型未考虑用户交互以及时间属性的影响，所以融合了时间属性以及用户之间交互特征，同时结合微博网络结构提出了 MUR 算法；最后，将算法与经典 PageRank 算法和 TS 算法做了比较。实验结果表明，模型在微博话题用户的时效性、话题推动以及对粉丝的影响力等方面表现较好，证明了模型的合理性和有效性。

关键词：关键用户；MUR；时间属性；用户交互

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.12.0795

Key users mining in micro-blogging topic based on user attributes

Ke Yang, Sui Jie†

(*School of Engineering Science*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Considering the timeliness of the microblogging topic and the feature of interaction between the users, therefore, this paper put forward a key user's mining algorithm based on user interaction to effectively find topic-sensitive key users. Firstly, this paper introduced the definition of key users in microblogging topic and its relevant characteristics. Secondly, in that the traditional models ignored the influence of user interaction and time attribute, this model fused the time property and the characteristics of interaction between the user together in the model at the first time, and the MUR algorithm was put forward with the combination of the microblogging network structure. Finally, it compared the algorithm with the classical PageRank algorithm and TS algorithm. The experimental results show that the model is more reasonable in terms of timeliness and topic driving, certificating the rationality and validity of the model.

Key words: key user; MUR; time; user interaction

0　引言

影响力通常被认为是可以改变他人思想和行为的能力。在传播学理论中，有这样一种人被称为意见领袖，在社交网络中经常向公众提供意见和建议。他们向受众传播信息，形成信息的二级传递。影响力存在于生活的方方面面，其中一个例子就是它经常应用于商业领域。因此，营销人员可以通过口口相传去建立大规模的连锁反映，从而更好地销售产品。

微博是基于社交关系来进行信息传播的媒体平台，作为重要的社交网站，微博引发了众多的关注和研究[1]。在微博中，只有很少的人才能成为话题的引领者。统计显示，约有 20%的微博用户发布超过 76%的微博。因此，只有少数用户在微博中特别突出，他们发布了大量微博，相比之下，大量用户充当了微博话题的接收者。微博中的用户影响力是通过影响他人的思想感情、激励他人产生更多的交互行为来体现的。

国内外研究人员对关键用户的挖掘方法也是层出不穷。相关学者提出了很多影响力度量方法[2~4]，通过比较入度、回复、转发、提及、读者数量等指标度量用户的影响力。其中，Bakshy 等人[5]利用消息扩散的结构信息预测和计算用户的影响力，大量实验数据表明 Twitter 上的很多热门话题都是由高影响力用户发起或传播。赵之滢等人[6]发现社团结构对于理解和控制消息传播有重要意义，并提出了一种基于网络社团结构的节点影响力度量方法，该方法能够更加真实地挖掘社会关系网络中对传播过程至关重要的 Hub 节点。毛佳昕等人[7]提出了一个通过预测用户传播信息能力大小来分析和度量用户社会影响力的方法，该方法将用户访问微博的时间以及用户偏好等行为因素与社会网络结构信息相结合，从而获得更好的影响力估计结果。

在基于给定话题的影响力计算方面，Tang 等人[8]提出了话题因子图模型同时支持同质和异质网络环境下的影响力分析。Weng 等人[9]提出了一个 TwitterRank 算法，该算法将话题相似

收稿日期：2017-12-05；**修回日期：**2018-01-25　**基金项目：**国家重点研发计划项目（2017YFB0803001）；国家自然科学基金面上项目（61572459）

作者简介：柯阳（1992-），男，安徽黄山人，硕士研究生，主要研究方向为数据挖掘、机器学习；隋杰（1976-），女（通信作者），山东日照人，副教授，博士，主要研究方向为 Web 挖掘，应急管理（suijie@ucas.ac.cn）．

度融合到基于网络结构的分析方法中，能够对不同话题领域的影响力用户进行排序。Saez-Trumper 等人[10]将时间属性融合到用户关系图中，并提出了一个基于 PageRank 的排序方法对特定话题的用户影响力进行排序。但是话题传播扩散过程中用户影响力是一个快速增长和衰减的过程，因此需要融合动态属性计算话题传播过程中的用户影响力。

从用户的个人属性来看，Cha 等人[2]认为高入度并非是高影响力的必要因素。他认为真正有影响力的用户在任何话题上都具有很高的影响力，通过研究入度影响、转推影响以及提及影响，他发现这三者之间并没有必要的联系，从而得出了真正有影响力的用户在任何话题上都具有很高的影响力的结论。Pal 等人[11]通过研究一系列的属性特征去找寻微博中的权威用户，通过构建用户影响力评价模型，最终产生给定话题的权威用户集。

在基于社会网络的信息传播方面，由于社会网络融合为社会计算等各项研究提供更充分的用户行为数据和更完整的网络结构,从而更有利于人们通过社会网络理解和挖掘人类社会,具有重要的理论价值和实践意义[12]。很多学者针对社会网络信息传播已经开展一些有意义的研究，提出了一些知名的信息传播模型包括独立级联模型[13]和线性阈值模型[14]。Sait 等人[15]针对社会网络中如何选择对应的模型做了深入的研究，并在独立级联模型和线性阈值模型基础上提出了改进方案，但是并没有考虑用户之间的交互对信息传播的影响，消息传播是一个动态的过程，需要通过传播模型动态的描述消息传播过程。Manue 等人[16]针对博客和新闻文章的传播研究跟踪传播路径和在网络中的影响，提出了一种近似的算法。但是作者为了研究简单，设定的前提假设网络是 静态且不变的，而微博网络中的信息是瞬变的。Fan 等人[17]针对新浪微博的消息扩散方式进行研究，发现热门事件的消息扩散拓扑结构呈星型或者两级结构。微博中的用户行为对消息传播起着重要作用，对关注和转发行为进行了广泛的研究[18, 19]。Michael 等人[20]研究关注和转发链接的语义表明主题的相关性能够很好的通过转发传递，通过关注和转发建立起 Twitter 图进行分析，但是并没有考虑交互双方的用户属性和时间因素。

通过对相关文献的调研发现，很多对关键用户的研究和方法主要集中于静态拓扑或者考虑单一属性。与之不同，本文创新性地在关键用户挖掘模型中引入用户交互属性和时间属性，提出了 Microblog User Rank(*MUR*)方法。实验结果表明该方法能够比较好地挖掘特定话题下的关键用户。

1　关键用户挖掘算法

在本章主要讨论三个部分：第一个部分讨论了对关键用户的定义；第二个部分讨论了微博话题关键用户的相关属性；第三个部分基于 PageRank 算法考虑结合了用户的相关属性提出了 *MUR* 挖掘模型。

1.1　关键用户的定义

在本文中把关键用户定义为比较早的接受话题趋势并激发更多交互使话题变得流行的人。作为一名“创新者”，关键用户必须比较早的接受话题趋势。对于关键用户来说，仅仅比较早的知道话题是不够的，因为其他微博受众接受话题可能与他们无关。对于一个特定的话题，用户交互在信息传播中扮演了非常重要的角色。关键用户应该是保持较高交互的一类人，他们通过产生更多交互来使话题变得流行。

1.2　关键用户的特征属性

1.2.1 时间属性

其他度量如入度甚至经典的 PageRank 算法都忽略了与话题相关的时间属性。这可能导致用户事实上根本没受目标用户的影响，而在所提的模型中却错误地计算了这个影响。不可以认为在话题已经流行之后才接受微博话题的用户是推动话题的关键用户，即便他拥有数以亿计的粉丝。

1.2.2 用户交互

在本文把关键用户定义为比较早的接受话题趋势并激发更多交互使话题变的流行的人。作为一名“创新者”，关键用户必须比较早的接受话题趋势。对于关键用户来说，仅仅比较早的知道话题是不够的，因为其他微博受众接受话题可能与他们无关。对于一个特定的话题，用户交互在信息传播中扮演了非常重要的角色。关键用户应该是保持较高交互的一类人，他们通过产生更多交互来使话题变得流行。在微博中，用户交互主要是通过用户之间的评论和转发体现。

1.3　关键用户挖掘模型

微博话题 k 可以利用特征形式化描述为 $\{f_1, f_2, \ldots, f_n\}$，其中 f_n 为微博话题 k 的第 n 个特征。 假设 $G_k(N_k, E_k)$ 表示微博话题 k 的用户关系图，其中 N_k 为微博话题 k 的用户集合，该用户集合内的每个用户至少发布了包含突发话题 k 的一个特征，E_k 表示边集合，对于任意的 $(u,v)\in E_k$，(u,v) 表示用户 u 与用户 v 之间的转发关系。为了挖掘微博话题演变过程中促使突发话题形成的关键用户，用户行为的时间属性显得尤为重要。本文将时间属性融合到用户关系图模型的建立过程中，对于用户 $v\in N_k$，令 $t_i(v)$ 表示用户 v 发布包含特征 f_i 的消息时间（如果用户 v 未发布包含特征 f_i 的消息，则 $t_i(v)=0$）。令 $s_1(v)$ 为用户–话题特征向量，该向量的每个分量 $s_1(v)_i$ 表示用户 v 是否发布描述特定话题 k 的特征 f_i 的权重，其计算公式如下：

$$s_1(v)_i = \begin{cases} 1, \text{if } t_i(v) > 0 \\ 0, \text{otherwise} \end{cases} \tag{1}$$

当用户 v 发布了微博话题 k 的特征 f_i 时，把 $s_1(v)_i$ 的值设置为 1；如果没有，则设置为 0。

定义 $I_{(u,v)_i}$ 为 u 对 v 关于话题k的特征f_i的交互权重序列：

$$I_{(u,v)_i} = [I_{(u,v)}(0)_i, I_{(u,v)}(1)_i, \ldots, I_{(u,v)}(T)_i],$$

其中：T 代表时间窗口的序号；$I_{(u,v)}(T)_i$ 表示在第 T 个时间窗口；u 和v关于话题k的特征f_i的交互权值大小。本实验中，时间窗口的长度设置为 1 天。在微博话题中，用户之间的交互主要

体现在评论与转发，所以定义 $I_{(u,v)}(T)_i$ 如下：

$$I_{(u,v)}(T)_i = \alpha \frac{Repost_{(u,v)}(T)_i + 1}{Repost_u(T)_i + 24} + \beta \frac{Comment_{(u,v)}(T)_i + 1}{Comment_u(T)_i + 24}, \quad (2)$$

其中：$\alpha + \beta = 1$, 。

这里 $Repost_{(u,v)}(T)_i$ 表示在第 T 个时间窗口，用户 u 转发 v 关于微博话题 k 的特征 f_i 微博的数量；$Repost_u(T)_i$ 表示全部用户 u 转发关于微博话题 k 的特征 f_i 微博的数量。同理，$Comment_{(u,v)}(T)_i$ 表示在第 T 个时间窗口，用户 u 评论 v 关于微博话题 k 的特征 f_i 微博的数量；$Comment_u(T)_i$ 表示全部用户 u 评论关于微博话题 k 的特征 f_i 微博的数量。如果用户 u 在第 T 个时间窗口既没有评论也没有转发，不能简单地把 $I_{(u,v)}(T)_i$ 设置为 0。在这里作了 Laplacian 平滑处理，得到平滑后的交互权值 $I_{(u,v)}(T)_i$ 。

令 $s_2(u,v)$ 为用户关系向量，该向量的每个分量 $s_2(u,v)_i$ 表示用户 u 与用户 v 针对特定话题 k 的特征 f_i 的边权重，其计算公式如下：

$$(u,v)_i = \begin{cases} \dfrac{\sum_{T=T_{t_i(v)}}^{T_{t_i(u)}} I_{(u,v)}(T)_i}{T_{t_i(u)} - T_{t_i(v)} + 1}, & \Delta > 0, \\ 0, & otherwise \end{cases} \quad (3)$$

其中：$\Delta = t_i(u) - t_i(v)$表示用户 u 与用户 v 发布特征f_i的微博时间间隔。$T_{t_i(u)}$ 表示 $t_i(u)$ 所在的时间窗口的序号数。当$\Delta \le 0$，即 $t_i(u) \le t_i(v)$，用户 u 发布的特征 f_i 的话题微博时间要早于用户 v，所以不能认为用户 v 向用户 u 传递了话题趋势。因此将 $s_2(u,v)_i$ 的值设置为 0。当$\Delta > 0$时，$s_2(u,v)_i$ 的值通过计算 $T_{t_i(v)}$ 到 $T_{t_i(u)}$ 这段时间 u 和 v 的交互均值得出。

对于任意的$(u,v) \epsilon E_k$，令 $I_k^*(u,v)$ 表示用户 u 对用户 v 的影响力，其计算公式如下：

$$I_k^*(u,v) = \left(\frac{s_1(v) \cdot s_2(u,v)}{\|s_1(v)\| \times \|s_2(u,v)\|} \right), \quad (4)$$

基于 PageRank 算法将 $I_k^*(u,v)$ 规范化公式如下：

$$I_k(u,v) = \frac{I_k^*(u,v)}{\sum_{\omega \in Out_{G_k(u)}} I_k^*(u,\omega)}, \quad (5)$$

令 $MUR_k(v)$ 为微博话题 k 的关键用户权重，其计算公式如下：

$$MUR_k(v) = (1-d) D_k(v) + d \sum_{\omega \in In_{G_k(v)}} MUR_k(\omega) I_k(\omega, v), \quad (6)$$

其中：$d (0 \le d \le 1)$为调节因子；D_k 表示用户在用户关系图上的概率分布。由于话题间是相互独立的，对于任意的$v \in N_k$，$D_k(v)$ 的计算公式如下：

$$D_k(v) = 1 / |N_k|, \quad (7)$$

为了算法能够收敛，这里将 d 的值设置为 0.8。用户的 MUR 值越大，即为促使微博话题形成的关键用户。

算法的过程见算法 1。通过输入开始节点的 ID、目的节点的 ID 以及边权重 $I_k^*(u,v)$，最终可以得到 MUR 值以及它们的节点 ID。

算法 1　MUR 算法表示

算法：*MUR*

输入：“start_id-des_id- $I_k^*(u,v)$ ”，迭代次数 N，阻尼系数 d

输出：排名-节点-权值

```
//迭代 N 次直到算法收敛
for iteration in range(0,N):
foreach node in node_set:
    if iteration == 0:
        //首次迭代所有节点初值设为 1
        node.set_value(1)
    else:
        //计算节点的入边影响力总和
        in_sum = 0
        foreach in_edge in node.in_edges:
            in_sum = in_sum + in_edge.value
        end foreach
        //计算当前节点的 MUR 值
        MUR = (1-d) + d * in_sum
        //保存当前节点的 MUR 值
        node.set_value(MUR)
        //计算节点出边权值总和
        out_sum = 0
        foreach out_edge in node.out_edges:
            out_sum = out_sum + out_edge.value
        end foreach
        //更新每条边的权值
        foreach out_edge in node.out_edges:
            new_edge_value=node.value*out_edge_value/out_sum
           //赋予每条边新值
            out_edge.set_value(new_edge_value)
        end foreach
    end if
end foreach node
end for iteration
```

2　实验方案及分析

所有的实验都是在 Intel 酷睿 I7-6700UCPU，8GB 内存，64 位操作系统上实现。Graphchi 是由 CMU(卡内基梅隆大学)博士 Aapo Kyrola[21]开发的一套基于磁盘的图处理系统，该系统能有效处理边数目达数十亿规模的数据图。在本实验室的硬件条件下，基于 Graphchi 框架[22]开发出关键用户的挖掘模型。

2.1　实验数据

SG YYGY 收集了腾讯微博 2016 年 1 月 1 日到 2016 年 12 月 31 日的 3 个话题共计 550 万条数据，话题涉猎政治/时尚和美食方面。同时，有超过 100 万微博用户以及他们之间的社交关系，详细描写如表 1 所示。

表 1 实验数据集描述

Topics	用户	原创数	转发数
#两会#	384671	184833	1102337
#VGirl#	646322	494093	2121899
#家乡美食#	318218	305637	1525895

为了下面描述的方便，用话题#1#、#2#、#3#来分别表示上面三个话题。

2.2 实验评价

为了验证本文算法的有效性，选择了一些经典模型进行结果比较。相应模型描述如下：

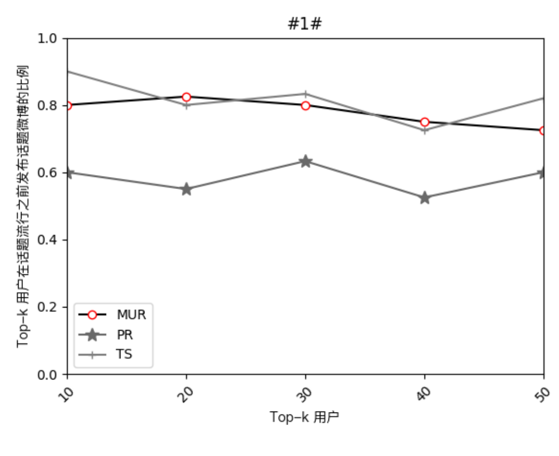
PageRank 算法：它是最流行的链接分析算法，用于对搜索引擎在用户查询后返回的结果进行排序。在微博社区中，如果一个节点被许多其他节点关注，那么它的 PageRank 值相对较高。然而当一个被大量节点关注的用户接受的话题微博在话题流行趋势之后的话，当然只能认为他是信息的被动接受者，而不能是推动话题流行的关键用户，所以这种方法挖掘出来的结果可能会不准确。

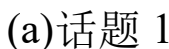
TS 算法[10]：它是基于 *PageRank* 将时间属性融合到用户关系图的排序方法。与前者不同，它考虑了时间属性，但是计算用户关系向量仅仅取决用户发布的微博时间间隔对于挖掘关键用户并不精确，无法得出其他人接受话题与所挖掘的关键用户之间是否存在相关关系。

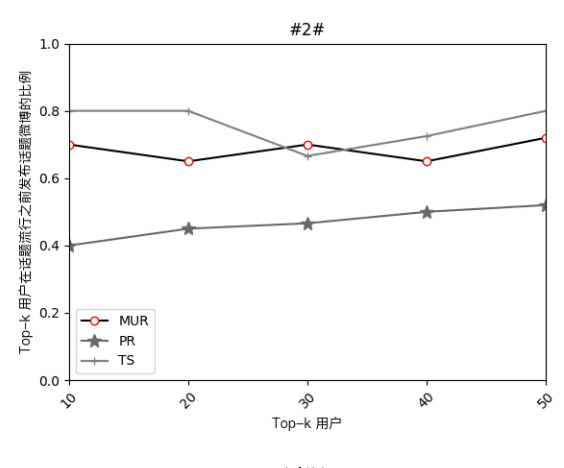
2.2.1 时间考量

时间来说对关键用户来说非常重要，因为挖掘一个在话题流行趋势之后才接受话题的“关键用户”是毫无意义的。全国两会是 2016 年一个非常重要的话题。根据话题的特征化定义为$k_{NPC\&CPPCC} = [\#两会\#, \#微博看两会\#, \ldots]$，共计 11 个话题特征。经过统计，网络图共包括 38 4671 个节点以及 2 443 924 条有向边。全国两会这个话题讨论的峰值出现在 2016 年 3 月 6 日，共计 79 443 个用户发布 118 786 条微博。根据上述提出的算法，本文计算网络图中至少发布一个话题特征的关键用户权值，与 *PageRank* 算法和 *TS* 算法进行比较，发现经典 *PageRank* 算法易于给一些过时的用户很高的权值，而 *TS* 算法和本文的 *MUR* 算法由于都考虑了时间属性，所以表现较好。为了说明这个，本文计算了 3 种算法 top- k (k =10,20,30,40，50)用户在峰值之前发布话题微博的比例。结果如图 1 所示。

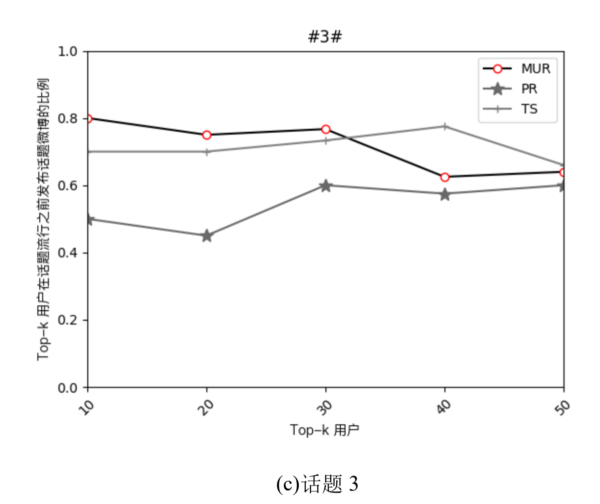
由图 1 可知，在任一话题 *TS* 算法和 *MUR* 算法都比 *PageRank* 算法表现要好。本文将其归因于 *PageRank* 算法由于不考虑时间属性更容易挖掘一些过时的人。尤其在关于时尚的#2#话题，*PageRank* 算法更是表现不佳。由于 *MUR* 算法和 *TS* 算法都考虑了时间属性，导致实验结果看起来好像没有太大差别。但是为了说明满足时间条件的 *TS* 挖掘模型挖掘出来的用户并非促使特定话题形成的关键用户，在接下来的部分做了些实验来进一步说明。




(a)话题 1



(b)话题 2



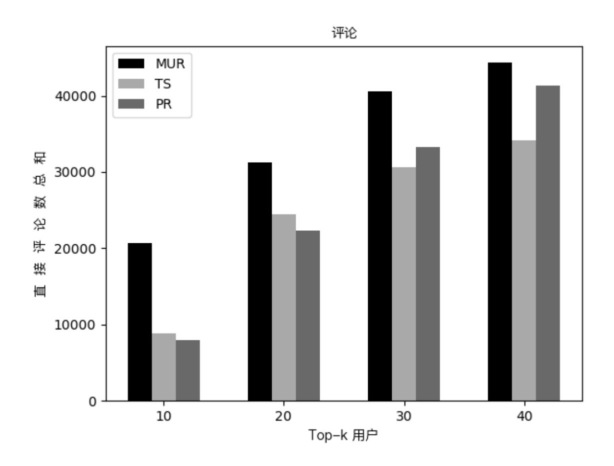
(c)话题 3

图 1 不同算法 Top-k 用户中在峰值之前接受话题的比例

2.2.2 用户交互

时间来说对关键用户来说非常重要，因为挖掘一个在话题流行趋势之后才接受话题的“关键用户”是毫无意义的。在微博中，用户话题影响主要通过他的粉丝的评论和转发向网络图中进行扩散。评论可以影响信息传播的深度，转发可以影响信息传播的广度。本文中认为用户微博被评论数和被转发数一定程度上能反映关键用户的影响力，所以计算了 top- k (k =10,20,30,40)用户微博直接被评论量和被转发量。实验结果展

现如图 2 所示。



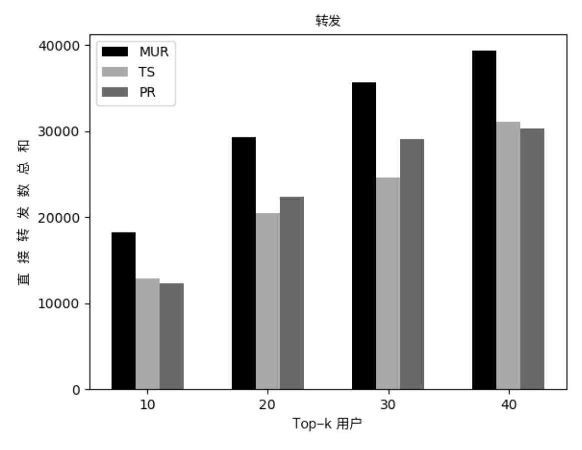


图 2　不同算法间 Top-k 用户直接评论量总和与直接转发量的总和

如图 2 所示，相比较其他两个算法，发现用户之间的交互在 *MUR* 算法上得到了更大的体现。很明显，MUR 挖掘的用户在考虑交互的情况下对网络中的舆论产生更加积极的影响，促使更多用户参与评论与转发。同时想知道满足时间条件的 *TS* 挖掘模型挖掘出来的关键用户是否促使了微博特定话题形成。在 top-20 的用户中，从影响信息传播广度的转发数上来看，*TS* 算法的总数比 *MUR* 算法少，甚至比 *PageRank* 算法还低，把原因归结于 *TS* 算法考虑的网络图中边的权值仅仅取决于时间间隔。所以只能认为 *TS* 算法挖掘出来的用户看起来更像个"创新者"，而不能被认为是关键用户。限于篇幅，本文在这里仅仅展示了话题 1 用户被评论与被转发的情况，在另外两个话题中，*MUR* 算法依然比其他两个算法表现更好。

本文算法致力于发现网络图中的一些节点，这些节点比较早得接受话题，而且有能力影响其他节点产生更多交互，最终使话题流行。为了进一步说明这个问题，本文计算了 top-10 的用户话题微博的全局转发量。实验结果如图 3 所示。

如图 3 所示，在涉及政治、时尚和美食领域的话题时，*MUR* 值高的 top 用户可以通过影响他们的粉丝产生更多的用户交互，从而推动话题的形成与发展。图 3 也意味着更多的人接受话题是因为 *MUR* 算法挖掘的 top 用户。更多的粉丝通过转发去扩散 top 用户的影响，最终推动了微博话题的形成。

2.2.3 社会关系影响力

令 $IFR_k(v)$ 去衡量 top 用户影响其社会关系的能力。其公式计算如下：

$$IFR_k(v) = {NIF_k(v)}\Big/{TNF(v)} \tag{8}$$

其中：$NIF_k(v)$ 表示 top 用户的粉丝中发布微博时间晚于 top 用户的数量；$TNF(v)$ 表示这些用户的全体粉丝数。

Top-50 用户平均影响粉丝率如表 2 所示。

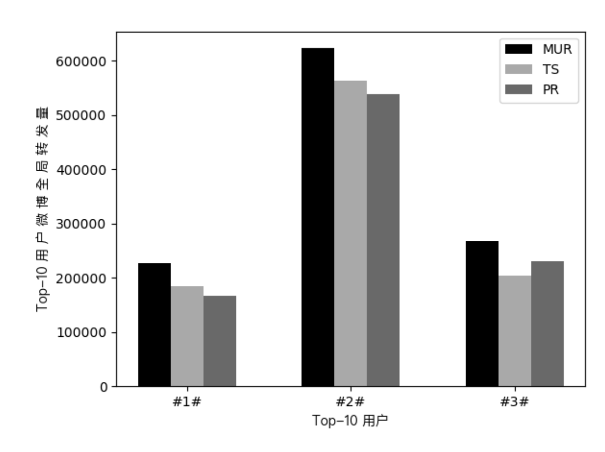


图 3　不同算法 Top-10 用户在微博网络全局转发量的总和

表 2　top-50 用户平均影响粉丝率

Topics	MUR(%)	TS(%)	PR(%)
#1#	0.058	0.050	0.039
#2#	0.134	0.113	0.105
#3#	0.085	0.081	0.070

从表 2 中可以知道，*MUR* 算法的 top 用户的 *IFR* 值比 *PageRank* 算法以及 *TS* 算法都要高，将其归因于 *PageRank* 算法挖掘了"过时"的用户和 *TS* 算法忽略了用户交互的影响。这也表明了高交互性的 *MUR* 用户更能影响其社会关系，从另一个方面也表明本文考虑关键用户属性的正确性。

2.2.4 算法相关性

本文尝试研究 *MUR*、*PageRank*、*TS* 这三种算法产生的关键用户列表的相关性。Kendall's τ是数学统计中一个常用的系数。τ的取值区间是[-1, 1]。如果两个用户列表一致，则$\tau = 1$；两个互逆序列的 Kendall's τ系数为-1。τ的值越大，表明更大的正相关性。

表 3　不同算法间的相关性

	#1#	#2#	#3#
MUR vs. TS	0. 4286	0. 4017	0. 4501
MUR vs. PR	0. 5187	0. 4984	0. 5327

表 3 展现了 *MUR* 算法与 *PageRank* 算法及 *TS* 算法之间挖掘用户之间的 τ. 值以发现本文算法与其他两种算法的结果并不是非常相似，这也表明本文算法在挖掘关键用户方面的实验

结果与其他算法相比有一定的区分度和不可替代性。

3　结束语

本文定义的关键用户是比较早的接受话题趋势并激发更多交互使话题变得流行的人。通过将时间属性和用户交互结合，提出了一种有效的算法去挖掘微博话题的关键用户。本文做了一系列的实验去验证本文模型，实验结果表明，本文算法在时间性和用户影响角度优于先前算法。考虑到用户交互存在其他的形式，未来计划进一步研究交互作用，从而建立一个更加准确的模型。

参考文献：

[1] 熊才伟，曹亚男．基于发文内容的微博用户兴趣挖掘方法研究 [J]．计算机应用研究, 2018, 35 (6): 1-6.

[2] Cha M, Haddadi H, Benevenuto F, et al, Measuring user influence in Twitter: the million follower fallacy [C]// Proc of International Conference on Weblogs and Social Media. 2010.

[3] Lee C, Kwak H, Park H, et al, Finding influentials based on the temporal order of information adoption in twitter [C]// Proc of International Conference on World Wide Web. 2010: 1137-1138.

[4] Ye S, Wu F. Measuring message propagation and social influence on Twitter [J]. Communication Networks and Distributed Systems, 2013. 11 (1): 59-76.

[5] Bakshy E, Hofman J M, Mason W A, *et al.* Everyone's an influencer: quantifying influence on twitter [C]// Proc of the 4th ACM International Conference on Web Search and Data Mining. 2011: 65-74.

[6] 赵之滢，于海，朱志良，等．基于网络社团结构的节点传播影响力分析 [J]．计算机学报, 2014, 37 (4): 753-766.

[7] 毛佳昕，刘奕群，张敏，等．基于用户行为的微博用户社会影响力分析 [J]．计算机学报 2014, 37 (4): 791-800.

[8] Tang J, Sun J, Wang C, *et al.* Social influence analysis in large-scale networks [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 807-816.

[9] Weng J, Lim E, Jiang J, *et al.* Twitterrank: finding topic-sensitive influential twitterers [C]// Proc of the 3rd ACM International Conference on Web Search and Data Mining. 2010: 261-270.

[10] Saez-Trumper D, Comarela G, Almeida V. Finding trendsetters in information networks [C]// Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 1014-1022.

[11] Pal A, Counts S. Identifying topical authorities in microblogs [C]// Proc of ACM International Conference on Web Search and Data Mining. 2011: 45-54.

[12] 周小平，梁循，赵吉超，等．面向社会网络融合的关联用户挖掘方法综述 [J]．软件学报, 2017 28 (6): 1565-1583.

[13] Watts D J. A simple model of global cascades on random networks [C]// Proc of National Academy of Sciences. 2002: 5766-5771.

[14] Goldenberg J, Libai B, Muller E. Talk of the network: a complex systems look at the underlying process of word-of-mouth [J]. Marketing Letters, 200112 (3): 211-223.

[15] Saito K, Kimura M, Ohara K, *et al.* Selecting information diffusion models over social networks for behavioral analysis [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. 2010: 180-195.

[16] Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence [C]// Proc of ACM Knowledge Discovery & Data Mining. 2011: 1019-1028.

[17] Fan P, Li P, Jiang Z, *et al.* Measurement and analysis of topology and information propagation on Sina-Microblog [C]// Proc of IEEE International Conference on Intelligence and Security Informatics. 2011: 396-401.

[18] Ghosh R, Surachawala T, Lerman K. Entropy-based classi cation of"retweeting"activity on Twitter [C]// Proc of KDD Workshop on Social Network Analysis. 2011.

[19] Hong L, Dan O, Davison B D. Predicting popular messages in Twitter [C]// Proc of International Conference on World Wide Web. 2011: 57-58.

[20] Welch M J, Schonfeld U, He D, *et al.* Topical semantics of twitter links [C]// Proc of ACM International Conference on Web Search and Data Mining. 2011: 327-336.

[21] Kyrola A. DrunkardMob: billions of random walks on just a PC [C]// Proc of ACM Conference on Recommender Systems. 2013: 257-264.

[22] Kyrola A, Blelloch G, Guestrin C. GraphChi: large-scale graph computation on just a PC [C]// Proc of Usenix Conference on Operating Systems Design and Implementation. 2012: 31-46.